Capturing social cues with imaging glasses

Capturing Social Cues With Imaging Glasses

Lauren Murray
University of St Andrews
St Andrews, KY16 9SX, UK
lm225@st-andrews.ac.uk

Ross Goucher
University of Stirling
Stirling, FK9 4LA, UK
ross.goutcher@stir.ac.uk

Philip Hands
University of Edinbrugh
Edinburgh, EH9 3FF, UK
philip.hands@ed.ac.uk

Juan Ye [a]
University of St Andrews
St Andrews, KY16 9SX, UK
juan.ye@st-andrews.ac.uk

[a]Correspondence author

UbiComp/ISWC '16 Adjunct, September 12-16, 2016, Heidelberg, Germany
ACM 978-1-4503-4462-3/16/09.
http://dx.doi.org/10.1145/2968219.2968260

Abstract
Capturing visual social cues in social conversations can prove a difficult task for visually impaired people. Their lack of ability to see facial expressions and body postures expressed by their conversation partners can lead them to misunderstand or misjudge the social situations. This paper presents a system that infers social cues from streaming video recorded by a pair of imaging glasses and feedbacks the inferred social cues to the users. We have implemented the prototype and evaluated the effectiveness and usefulness of the system in real-world conversation situations.

Author Keywords
Affective Computing; Imaging glasses; Emotion Recognition

ACM Classification Keywords
H.5.m. [Information Interfaces and Presentation (e.g. HCI)]: Miscellaneous

General Terms
Design, Algorithms, Experimentation, Measurement, and Performance

Introduction
Social interaction refers to the way in which people interact with one another. This is an important aspect of our everyday lives since communication with others is the main con-

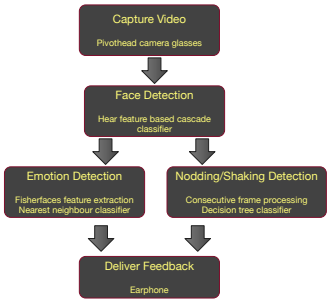

Figure 1: The system has three main components: streaming video via imaging glasses, inferring emotion valence and agreement in a reasoning engine, and delivering feedback via earphone.

Table 1: Comparison of emotion detection performance with different feature extraction technique and different facial areas. The Fisherface with the mouth area performs best.

Feature	Face	Eyes	Mouth
Eigenface	0.24	0.31	0.67
Fisherface	0.24	0.3	**0.73**

tributor for building both personal and professional relationships. Many facets of communication are nonverbal, such as eye contact, facial expressions, and body postures, all of which are useful for social perception [2]. However, this can leave visually impaired people at a disadvantage in social situations. The inability to see this type of communication can lead blind people to miss these social cues, which can make them feel uncomfortable and may even discourage them from further engaging with others [5].

Assistive technologies have been designed to help deliver social cues to visually impaired people [4]. McDaniel et al. [5] have designed a haptic belt that use vibration location and duration to communicate the direction and distance of the conversation partner, via the computer vision algorithms. Krishna et al. [3] have designed a haptic device, called VibroGlove, to convey facial expressions via various vibration patterns that symbolise the emotional icons. We propose a system that infers and delivers high-level social cues from real-time streaming video. Here we consider the social cues in two dimensions: *emotion valence* (i.e., positive and negative emotions) and *agreement* (i.e., nodding and shaking). We have designed user studies to evaluate the effectiveness of the system in real-world environments. The initial evaluation result is promising, even though we have encountered problems when stimulating and evaluating spontaneous emotional expressions. We will share the lessons that we have learned.

System Design

The system design is shown in Figure 1. We use the the Pivothead camera glasses[1] to capture live video, which are output as an rtsp stream and accessed via a URL. Ffmpeg is used to split the live video into frames in real time. We use the state-of-the-art computer vision algorithms to process facial images, detect faces, and infer emotion valence. We use the Haar-like feature based cascade classifier [8, 9] to detect the face area in each frame. To infer emotion valence, we have experimented with different feature extraction algorithms like Eigenface [7] and Fisherface [1] in the bytefish framework developed by Philip Wagner[2], and also considered different areas of the face for emotion detection such as the whole face, eyes, and mouth. We use the nearest neighbour with the cosine distance as the classifier. We have evaluated different combinations of the strategies on the AT&T face image dataset [6], and in the end, the Fisherface feature extraction algorithm on the only mouth area performs best. The benchmark evaluation results are presented in Table 1.

To detect agreement, we extract head movements from a series of continuous frames; that is, calculating the x- and y-dimension differences of the face positions. Then a decision tree is used to classify whether the head movement indicates nodding or shaking. We have collected frames from a number of users performing staged nodding and shaking. We run the 10-fold cross validation on the dataset and the benchmark accuracies of detecting nodding and shaking are 70% and 81% respectively.

To note that we pre-train both the emotion valence and agreement detection algorithms on the AT&T facial dataset and our own collected dataset, and use them to infer in real-time videos.

Once both emotion valence and agreement have been inferred, we deliver the feedback via the earphone to the user; that is, we use the native speech software built in the Mac OS X operating system. The system says "positive"

[1] http://www.pivothead.com

[2] https://github.com/bytefish/facerec

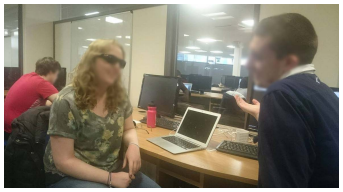

Figure 2: Two participants have a conversation in a lab environment, where the female participant wears the imaging glasses and receives the social cue feedback that infers the male participant's nonverbal behaviours.

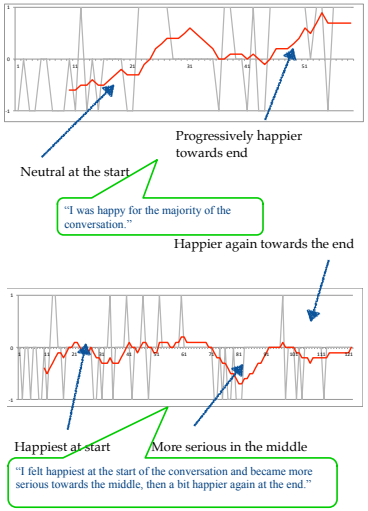

Figure 3: Emotion valence detection over the course of conversations

and "negative" for the inferred positive and negative emotions, and "agree" and "disagree" for the inferred nodding and shaking. To avoid interfere much with the conversation, we only deliver the feedback when the inferred result is different from the last one.

User Study and Evaluation

The main objective of the evaluation is to assess whether the system can accurately capture social cues and whether it can deliver these social cues that are useful for users to better understand the social situations. To do so, we have conducted a *conversation-driven in the wild* user study, which we will detail in the following.

Procedure 10 users have been recruited for this study. The participants are grouped into paris: one of them is blindfolded and wears a pair of Pivothead camera glasses, which are used to record the other participant during the conversation. The glasses wearer participant also wears an earphone in one ear to receive the feedback that delivers the inferred social cues. In the following of the paper, we refer the glasses wearer and the other participant as a target and partner participant respectively.

At the beginning of each experiment, a target participant is given a collection of topics that are supposed to stimulate certain emotions and then both participants are asked to freely converse for about 2 minutes. Figure 2 has shown an exemplar scenario.

During their conversation, the system will process the streaming video from the image glasses, infer the agreement and emotion valence, and feedback to the user via the earphone. At the end of the experiment, the partner participants are asked to state their emotions and agreement they have felt throughout the conversation. However, it turns out that the agreement or disagreement is much harder to specify and has not occurred very often in the conversations. The emotions and agreements stated by the partner participants are used as the ground truth for analysing the performance of the algorithms. The target participant is also asked to fill in a post-study questionnaire, which is used to evaluate the usefulness of the feedback provided by the system.

Result Here we mainly conduct the qualitative evaluation; that is, how the inferred emotional valence and agreement match the feelings reported by the partner participants. Figure 3 presents two of the results, where the inferred emotion valence for each frame (-1 for negative, 1 for positive, and 0 for neutral, and the inferred points are linked in a thin grey line) and the aggregated emotion for the previous ten inferences is denoted in a thick red line.

In terms of the top figure of Figure 3, the partner participant's feedback is "I was happy for the majority of the video", and the system has inferred 58% of the time when the positive emotion is inferred. However, there are quite a few negative emotions inferred at the start of the conversation, which is supposed to be neutral.

In terms of the bottom figure, the participant's feedback is "I felt a mixture of all three emotions, I was happiest at the start and then became progressively more negative towards the end." This figure has shown a variation of the emotions: more mixed inference for the first half of the conversation and the majority of negative emotion is inferred for the second half. Similarly, the negative emotions are incorrectly inferred at the start of the conversation. We have manually examined the videos and found out that the algorithms often mis-classify neutral or implicitly positive emotions (that is, there is subtle happy expressions on the face) for negative emotions. Also there often exists a gap between people's expressed and self-perceived emotions; for example,

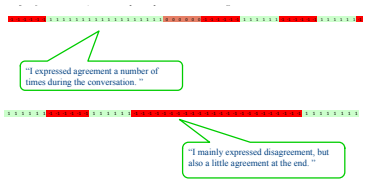

Figure 4: Agreement detection over the course of conversations

there are occasions where the partner participant stated happy, however we cannot tell it from their facial expressions in the video.

Figure 4 presents the results on inferring agreement and disagreement (1 indicates agreement, marked in light green, 0 indicates neutral, marked in light red, and -1 indicates disagreement, marked in dark red). Qualitatively the inference results roughly match the description of the participants: the first participant expressed "I expressed agreement a number of times during the conversation", while the second participant expressed "I mainly expressed disagreement, but also a little agreement at the end."

Detecting agreement and disagreement turns out to be a more tricky task than what we expect. First of all, the nodding and shaking detection can be affected by the head movement of the glass wearer; that is, if the target participant is moving their head, then the algorithm that only considers the change of the positions of the face positions in frames will infer nodding or shaking, even though the partner participant has not moved their head at all. This can be resolved by attaching an accelerometer on the glasses to detect the movement of the glasses wearer. The other issue is that shaking does not necessarily suggest disagreement, nor does nodding suggest agreement. For example, often when we agree with something negative, we tend to shake our head. It is context sensitive. In the post-questionnaires, the target participants report that they have felt a bit strange when the system told them "disagree", however they have had a strong feeling of their partners definitely agreed with them. We should have delivered the feedback simply as nodding or shaking, rather than translating them to agreement or disagreement.

Table 2 lists the questionnaires from the participants on whether the feedback is useful or intrusive. The majority of the target participants are positive about the system; that is, they consider most of the inferred social cues are consistent with their perceived feelings and as well as useful to help deal with uncertain situations and confirm their own decisions.

As we manually examine the video, we have found that the system performs better when the target participant does not do as much talking. The reason is that when they do lots of talking, their emotion is well expressed in their speech, and there are less uncertainties for the glass wearer participants to figure out.

Table 2: Questions and responses from the participants

Question	Answer
How consistent is the feedback with what you perceived your partner's emotions to be? (1 being not consistent, and 5 being very consistent)	The average rating is 3.
Does the feedback help your judgement of your partner's emotions? (Yes or No)	3 out of 5 answered "Yes".
Is the system intrusive? (Yes or No)	4 out of 5 answered "No".
Please explain how the feedback did or did not help you judge the emotions of your partner	The feedback was helpful as it confirmed what I already thought. The feedback was't helpful or unhelpful, as I didn't pay a lot of attention to it. The feedback wasn't helpful as there was never a point where I couldn't work out the emotion through the person's voice. The feedback helped confirm my judgement. The feedback helped a little when I wasn't sure how the person would feel in the situation they were describing.

Conclusion and Future Work

This paper presents a prototype system that infers and feedbacks social cues from streaming video in real time, which has a potential to help visually impaired people to understand social situations and be more confident in conver-

sations. We have designed and run proof-of-concept user studies, where participants are having free conversations in the real-world environment. The preliminary qualitative evaluation result is promising; that is, the algorithms can capture the emotional trend. However, there is quite a large space for future improvement.

We have not conducted proper quantitative evaluations on the accuracies of the algorithms, as manually examining and annotating each frame of all the videos is a very time- and effort-consuming task, and as well as has the risk of being subjective to individual interpretations. Our next step is to either get behaviour scientists to help annotate the videos or recruit multiple users to cross annotate the videos to balance the bias.

We will also work on the emotion valence and agreement detection algorithms. Currently, we mainly use the off-the-shelf algorithms, and as shown in the evaluation results, the accuracies are not particularly high at the frame-level, and they can be very fluctuating. Also the current algorithms are sensitive to noise; for example, when the target participants covers their faces with hands or an object, the algorithm cannot detect a face and thus fail to make a correct inference. In the future, we will look for ways to make the algorithms more robust and accurate.

In terms of feedback technologies, we are using the spoken sound as the feedback mechanism, which can interfere with the normal conversations. We will look for more subtle feedback and as well as how to make feedback easily comprehensible by users.

Last, an interesting observation rises from the current user study is how people integrate external sensory information with their own sensory perception; for example, integrating the inferred visual social cues with their own perceived verbal social cues. The external information can help people to make better decisions when they are uncertain. We can explore this further: how people trust the external information, or how the accuracy or confidence of external information affects the way people integrate them. This is a research topic relevant to human sensory perception and integration. We plan to work with psychologists on designing targeted user studies to pursue this problem.

REFERENCES

1. Peter N. Belhumeur, João P. Hespanha, and David J. Kriegman. 1997. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Trans. Pattern Anal. Mach. Intell.* 19, 7 (July 1997), 711–720.
2. Jon E. Grahe and Frank J. Bernieri. 1999. The Importance of Nonverbal Cues in Judging Rapport. *Journal of Nonverbal Behavior* 23, 4 (1999), 253–269.
3. Sreekar Krishna, Shantanu Bala, Troy McDaniel, Stephen McGuire, and Sethuraman Panchanathan. 2010. VibroGlove: An Assistive Technology Aid for Conveying Facial Expressions. In *Proceedings of CHI EA '10*. ACM, New York, NY, USA, 3637–3642.
4. Sreekar Krishna and Sethuraman Panchanathan. 2010. *Assistive Technologies as Effective Mediators in Interpersonal Social Interactions for Persons with Visual Disability*. Springer Berlin Heidelberg, Berlin, Heidelberg, 316–323.
5. T. McDaniel, S. Krishna, V. Balasubramanian, D. Colbry, and S. Panchanathan. 2008. Using a haptic belt to convey non-verbal communication cues during social interactions to individuals who are blind. In *Proceedings of HAVE 2008*. 13–18.
6. F. S. Samaria and A. C. Harter. 1994. Parameterisation of a stochastic model for human face identification. In *Proceedings of the Second IEEE Workshop on Applications of Computer Vision*. 138–142.
7. Matthew Turk and Alex Pentland. 1991. Eigenfaces for Recognition. *J. Cognitive Neuroscience* 3, 1 (Jan. 1991), 71–86.
8. P. Viola and M. Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of CVPR 2001*, Vol. 1. 511–518.
9. Phillip Ian Wilson and John Fernandez. 2006. Facial Feature Detection Using Haar Classifiers. *J. Comput. Sci. Coll.* 21, 4 (April 2006), 127–133.